NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

BC2406 - Analytics I

Seminar Group 8 Team 7
Submission Date: 6 November 2021

Team Members:

Oh Yee Hong, Ignatius (U2010058E)

Loi Xue Yin (U2010140K)

Lim Ruo Ting Bethany (U1910717C)

Joel Mui Kit (U2010877H)

Table of Contents

1. Executive Summary

In a fast paced world, telecommunications plays an integral role in the fundamental operations of the society. This report aims to solve a problem that many telecommunication companies face, specifically **M1** in our report, which is the loss of customers, also known as customer churn. In addition, to shed light on what might be the root cause of the problem and devise solutions to mitigate it.

The dataset used for this project is from the following link: https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113.

We started off by exploring the dataset to gain a better understanding of the variables. From our exploration, our group decided that **\`churn_value\`** was going to be the dependent variable. Prior to data exploration, we cleaned the data and preprocessed the data until it was appropriate.

From there, we conducted data exploration and visualisation to further explore the plausible relationships between the variables in the dataset, gathering valuable insights for our recommendations. All these were carried out with **ggplot2.**

Moving on from data exploration and visualisation, our group went on to analyse the data with the use of logistic regression and classification and regression tree. These two models helped us to identify variables that were statistically significant in the dataset, formulating the solutions proposed to reduce customer churn. To determine which model was suitable, our group decided to utilise the F1 Score to determine the better model.

The solutions proposed to M1 were aligned to our goal of this report: to reduce customer churn. To put it another way, our solutions were meant to increase the loyalty of the customers. These solutions included building a customer loyalty program and the enhancement of bundling of products and services. These will allow for M1 to reduce customer churn, increasing their revenue.

To sum up, as we apply our theoretical knowledge and concepts to this project, we hope to provide M1 with a comprehensive and dynamic analysis that is relevant to their business, ultimately reducing customer churn and seeing an increase in revenue.

2. Introduction to Analysis

2.1 Background Information

Singapore’s telecommunications (telco) industry was valued at 9.51 billion dollars in 2018. Telcos provide a host of services such as phone lines, mobile data and broadband internet to both private customers and businesses. With a mobile penetration rate of 150% and wireless broadband penetration rate of around 190% (*CNA, 2019*), Singapore is one of the most connected countries in the world. As the rollout of fifth generation (5G) mobile networks continues to pick up pace, we believe that the telecommunication companies will play an even larger role . (*CNBC, 2017*)

M1 is a mobile network operator (MNO) that provides full services communications in Singapore. Established in 1994, it is one of the original “Big 3” telcos in Singapore. However, the arrival of new mobile virtual network operators (MVNO) to the telco industry has caused the market to grow increasingly competitive. As of 2021, there are now nine MVNOs and four MNOs in Singapore. Starting with Circles.Life in 2016, MVNOs have disrupted the market by offering lucrative contract-free plans with abundant mobile data. The allure of these value-for-money data plans has convinced customers to cancel their existing plans and switch telcos. With the increasing competition in the industry, telcos will be looking to place more emphasis on customer retention.

Our team senses the opportunity to analyze customer data with the help of machine learning models and obtain valuable insights about the key factors that affect a customer’s decision to stick with their telco instead of changing to another.

2.2 Analytics Problem

As the telco industry in Singapore grows more saturated, telcos face the risk of losing their customers to competitors if they do not meet their needs. A study conducted by SMU about the info-comms industry revealed that of the four subsectors, the mobile telecom sub-sector had the biggest dip in customer satisfaction, falling 1.4% year-on-year.

Customer churn is a term that refers to the loss of clients. It is a key performance metric for companies in the telco industry, as a renowned study by Bain and Company found that a mere 5% increase in retention rates could increase profits by up to 95%. In addition, companies could save on marketing costs which is a huge operating expense, allowing for better profit margins. (*Beehive, 2020*) This provides a great incentive for telcos to understand the factors causing customers to leave and devise more effective customer retention schemes.

As we were unable to obtain customer data from Singaporean telcos, our team will be using a telco customer churn dataset provided by IBM to carry out the analysis. The dataset contains information about 7043 customers from a California-based telco in the USA. By using this dataset, we are making the assumption that the behaviour of customers in the US is similar to those in Singapore.

2.3 Planned Approach

Our approach to analyzing this problem can be broken down into three stages:

Stage 1: Data Cleaning

To ensure data is correct, consistent and usable, we will scan through the dataset to filter out irrelevant columns, remove duplicates and handle any missing data. We will have to ensure that there are no inconsistencies in our dataset such as typos or incorrect capitalisation as they can cause mislabelled categories or classes. Any errors will be reformatted accordingly. In addition, we have to decide how to tackle missing data as many algorithms will not accept missing values. There are several options to consider:

1. Drop observations that have missing values. However, if the quantity is large enough, doing so may result in loss of information that may have an impact on the results of further analysis.
2. Input missing values based on other observations. However, there is an opportunity to lose integrity of the data because we may be operating from assumptions and not actual observations.
3. Alter the way the data is used to effectively navigate null values

We will be using these options interchangeably, depending on the significance of the missing data.

Lastly, as the survey data mainly consists of strings, it needs to be encoded into nominal/ordinal categorical variables so that further analysis can be done.

Stage 2: Data Visualization

Using ggplot2 and other visualization packages, we will create charts and plots to give an overview of the relationships between different variables in the dataset.

Stage 3: Data Analysis with Machine Learning Models

Our team will utilize two methods of machine learning to aid our analysis:

1. Logistic Regression
2. Classification and Regression Trees (CART)

Logistic regression could be employed to predict categorical Y variables. The models will be able to help us identify the relationships between our variables available in the dataset via the use of odds ratio. Lastly, We will test the accuracy of the model by using a confusion matrix.

Decision trees could be used to help make sense of customer survey results and classify customers into different segments based on different traits. Identifying these business insights are highly valuable as they could assist in developing customer retention plans or highlight key aspects of the customer experience that contribute to sales. The accuracy of our predictive model will be determined based on node purity and gini index.

To obtain the desired results, we have identified variables of key importance in the dataset. For example, the churn score and churn value of the dataset. We will build decision tree models to uncover the factors that have the most significant influence on customers giving the desired response, using this information to develop tailored business strategies.

3. Data Cleaning and Preprocessing

3.1 Initial State of Data

For this analysis, we are using one dataset about telco customer churn. Before preprocessing, it contains **rows and 50 columns**. The dataset includes some basic patient traits as well as medical factors, such as type of time in hospital, number of medications etc.. Data can be classified into categorical variables (binary and polytomous) and continuous variables.

3.2 Removal of Irrelevant Columns

To begin the data cleaning process, it is important to remove columns that are irrelevant to the analysis.

1. Geographical data (\`Country\`, \`State\`, \`City\`, \`Zip Code\`, \`Lat Long\`, \`Latitude\`, \`Longitude\`)

As the focus of the analysis does not include the effect of geographical locations of customers on their likelihood to change telcos, any geographical data is irrelevant. Furthermore, as the data is collected from customers in California USA, geographical data is irrelevant for providing insights about the behaviour of Singaporean customers.

2. \`CustomerID\`

Customer ID is an unique identifier generated by the telco to keep track of its individual customers. Since the focus of the project is on identifying the factors that have the greatest impact on likelihood of customer churn, we have no use for identifying specific customers and should remove the column.

3. \`Count\`

Count is a column that was intended by the creator of the dataset to be used in dashboarding for counting the number of customers in a filtered set. It consists of a single value ‘1’ and has no analytical purpose, so we have removed it.

3.3 Dealing with Missing Values

Of the three aforementioned methods of dealing with missing values, dropping observations was chosen. Analysis of the dataset revealed that there were 11 observations with NAs in the dataset. Since the missing values were categorical and not continuous, we could not replace them with the mean of the column. As the number of observations with missing values makes up only 0.15% of the dataset, dropping them will not result in a loss of information that has a significant impact on the results of the analysis.

3.4 Encoding Categorical Variables

As mentioned earlier, there are two types of categorical variables: binary and polytomous. Binary categorical variables largely consist of 'Yes/No' responses and can be manually encoded with numerical values. Polytomous categorical variables require one-hot encoding, which we performed using the R package 'fastDummies'.

3.5 Replacing Incorrect Values

To separate the two types of categorical variables, we filtered the columns by the number of unique values: those with more than two were polytomous and those with two were binary. However, after completing the rest of the encoding process and testing the data with a CART model, we found that there was an error somewhere in the data table causing the model to present strange results.

After closer inspection of the data table, we found that there were seven columns that contained a third value, either 'No internet service' or 'No phone service'. It turns out that they were dependent on the `phone_service` and `internet_service` columns, which made sense logically. For example, if a customer does not have internet service in their contract with the telco, they would not be able to enjoy movie streaming services. These values should be changed to 'No' to standardize the data as it has the same meaning.

3.6 Renaming Columns

Before cleaning, every word in the dataset column is capitalized and there are spaces between each word. This could make code messier as backticks would have to be used every time a column needs to be referenced. To standardize the formatting, all spaces were replaced with underscores and column names were changed to lowercase.

3.7 Dealing with Data Imbalance

When doing some preliminary data exploration, we discovered that there was a significant imbalance in the distribution of the decision variable, ‘churn_value’.

Out of 7032 recorded instances, only 26.6% corresponded to having a churn value of 1 (Figure 1). As the dataset comprises all customers from the telco, it is not a result of sampling bias and suggests that the data is inherently skewed. One of the downsides of having data imbalance is that it affects the accuracy of machine learning algorithms and may lead to less accurate predictions (Brownlee, 2019).

From the multiple methods available to solve data imbalance, random oversampling appears to be most applicable to the dataset. Undersampling is likely inappropriate as the dataset only has 7032 rows, so randomly deleting data from the majority class could lead to insufficient data for machine learning models as they grow in accuracy the larger the dataset. Oversampling is a technique where examples from the minority class (churn value of 1) are duplicated at random, helping to rebalance the training dataset (Brownlee, 2020). However, it may cause the model to be overfitted to the dataset and affect predictive accuracy. As the ratio of imbalance is not too extreme, we have decided it would be best to just leave the data as it is.

4. Data Exploration & Visualization

4.1 Telco Customer Churn Dataset

The Telco Customer Churn Dataset consists of information about a Telco Company providing home phone and internet services to about 7000 customers in the third quarter of the year. There are various variables in the dataset which allows us to plot several graphs. However, we will mainly focus on the churned value which indicates if the customer has left the company in this quarter. After analysing what could have been the motivation behind the churned customers, we will provide recommendations specifically catered to the Telco company.

4.2 Demographic Categorical Variables

The dataset includes various demographic categorical variables such as a customer's gender, age, marital status and if they have any dependents.

As we can see from *Figure 2.1*, the dataset consists of a fair mix of both females and males, with a similar amount of churned customers for both genders as well. This allows us to be indifferent when analysing data with regard to genders.

A senior citizen is identified as a person who is 65 years old and older. A large majority belongs to people under 65 years old, which is more relevant when we are analysing and providing recommendations, because our target audience should belong to a younger group of people as they are generally perceived to be more tech-savvy.

Having a partner indicates if the customer is married. In this dataset, generally more customers without a partner will be more likely to churn compared to those with a partner. This could be attributed to the fact that most couples are engaged to the same Telco company, and it will be a hassle to switch from one Telco to another.

Last but not least, dependents refer to children, parents or grandparents. In this dataset, customers who live with dependents are very unlikely to churn. This could also be due to the fact that most families share the same Telco company, while those living without any dependents can make the decision to switch easily.

4.3 Distribution of Tenure/Tenure Months vs Churned Customers

Figure 2.2 measures the number of churned customers against their tenure months. The graph highlighted in blue shows the density of churned customers during the quarter while tenure months refers to the total amount of months that the customer has been with the Telco company by the end of the quarter. Observing the figure, it shows that a huge influx of customers tend to leave at the start of the tenure, accompanied by a gradual decrease.

Therefore, we can infer that the services provided by the Telco company might not meet their expectations, resulting in them switching to another Telco immediately. In addition, as a new Telco company, they might not have many brand-loyal customers, whereby customers are loyal to a specific brand or company.

Hence, this might be a critical factor in making customers stay for a long period of time. It is important to provide fast and quality internet services to customers, such that they will be motivated to continue using the services of the Telco company. In order to do so, there’s a need for the Telco company to constantly upgrade and invest in their services. Once customers are enticed to stay, it will enable the company to also build on their brand reputation in the long run and attract more potential customers.

4.4 Proportion of customers with different types of contract

We also considered the Contract variable which shows the Telco’s customers' current contract type. As seen in *Figure 2.3*, there are significantly more customers opting for month-to-month contracts instead, compared to the other 2 contract types.

This indicates that more customers are leaning towards short-term contracts, which ensures that they are not bound to long-term contracts and are able to terminate within a short period of time. There is also a trend showing the number of churned customers were more likely to be on a month-to-month contract basis, implying that more customers prefer not being tied down by certain contract terms. This further signifies a need to improve on their services as mentioned in *Figure 2.2* earlier.

In addition, it will be an advantage for the Telco company to revamp the terms and conditions for the month-to-month contracts, such as lowering charges in an attempt to attract more customers to sign up. On the other hand, it will also be beneficial to look into other existing contract plans to make it more appealing to customers, such that there are different contracts catered to different needs of customers.

4.5 Monthly Charges across Contracts

Monthly charges refers to a customer's current total monthly charge for all their services from the company. As seen in the chart in *Figure 2.4* for all three different types of contract, churned customers are always having significantly higher monthly charges as compared to retained customers. This might indicate that the high charges are a major reason why customers are opting to terminate and leave the Telco company.

Besides considering the attractiveness of the terms and conditions of the contract, it is important to take into account the affordability of the different contracts. The prices will also be an important factor that customers will give thought to in comparing between different Telco companies.

4.6 CLTV of customers with different types of contract

Customer lifetime value(CLTV) represents a customer's value to the Telco company over a period of time. It is being used to determine how much a customer is worth in comparison with others and hence also an important metric to help the Telco company develop strategies to retain existing ones.

Figure 2.5 shows the customer lifetime value of customers with different types of contracts. The blue line describes all customers who have churned, whereas the red line describes all customers who still remain with the Telco company. The graph shows a general upward trend of customer lifetime value across the tenure months under all three types of contract. The longer the tenure months, the higher the customer lifetime value.

One interesting finding illustrated in *Figure 2.5* is that the starting customer lifetime value of customers with month-to-month contracts is significantly higher compared to those with one-year and two-year contracts. Moreover, customers under a two-year contract who churned are the most valuable customers. Therefore, the Telco company should closely monitor these customers for churn and allocate resources for each class of customers accordingly.

4.7 Internet Service Subscription

The telco offers two types of internet connections, which are mainly fiber optic and DSL. Based on our observations, fiber optic internet service has a higher customer subscription compared to DSL. Customers' preference towards fiber optic over DSL could be attributed to the fact that they value speed and network reliability when choosing their internet service. This is because Fiber optic has a higher speed and is typically more reliable than DSL.

However, referring to *Figure 2.6(a)*, the proportion of customers who churned under the fiber optic internet service is significantly higher than those under DSL. One possible reason is that fiber optic internet service is more expensive. As shown in *Figure 2.6(b)*, the monthly charges of customers with fiber optic internet service are higher than those who subscribed to DSL and without internet service.

In order to reduce the number of churned customers, we will first have to ensure that the high price charged to customers who opted for fiber optic is justifiable. By providing them with higher speed and more reliable internet services, the high prices will be accounted for and in turn, help to retain customers in the long run. For customers who are deterred by the high charges, we can advise them to opt for DSL instead if high internet speed is not their main priority. This allows us to cater to different groups of customers who have different priorities when it comes to choosing a Telco Company.

4.8 Additional Services

The Telco company offers 6 additional services, including online security, online backup, device protection plan, premium technical support, streaming of tv as well as movies. From what we observed, all customers with no internet service subscription will not have all the add-on subscriptions. This indicates that all these 6 additional services are provided exclusively to those who subscribed to the internet service. The number of add-ons each customer opts for is entirely dependent on their preference.

As shown in *Figure 2.7*, there is a downward trend of customer churn when the number of additional services subscribed by customers increases. This trend suggested that customers who subscribed to more add-ons would have a lower churn rate. Thus, offering additional services to customers could make a great difference when it comes to giving customers reasons to stay.

Furthermore, customers who do not subscribe to any additional services have the highest proportion among those with internet service. If the Telco company is able to turn these groups of customers into a subscriber of add-on services, they would probably be able to reduce the churn rate of customers who subscribed to internet services overall.

5. Data Analysis with Machine Learning Models

5.1 Logistic Regression

Logistic Regression is a statistical model that is often used for predictive analytics and modelling by explaining the relationship between the categorical dependent variable and one or more independent variables by estimating the probabilities using a logistic regression equation.

To begin, we conducted a train-test split (2/3) so that we can train our model on the known dataset (train set) and subsequently test the model on ‘unseen’ data (test set).

Our dependent variable `churn_value`, is a categorical variable with a binary outcome: 0 for retained and 1 for churned. The initial logistic regression model, **LR1**, gave us 25 independent variables (*Figure 3.1*). From there, we identified the independent variables that were statistically significant, and built a second logistic regression equation, **LR2** (*Figure 3.2*)**.**

Our group used the ‘Performance’ package to verify whether the variables in **LR2** were collinear. The Variance Inflation Factor (VIF) was used to measure the magnitude of multicollinearity of model terms. A VIF less than 5 indicates a low correlation of that predictor with other predictors (rdrr.io, n.d.). From *Figure 3.3*, it shows that our variables have low correlation with one another and hence the low multicollinearity.

Following that, our group went on to calculate the odds ratio for each variable (*Figure 3.4*) and their respective confidence interval (*Figure 3.5*). As the odds ratio for **`contract_month_to_month`** has the highest factor value of 7.37, it shows that it would be the most important factor in our business to retain a customer. As the confidence interval levels exclude 1, the odds ratio is statistically significant.

We continued to use the ‘pROC’ package to determine the Receiver Operating Characteristic Curve (ROC) and Area Under Curve (AUC) of our trainset. ROC gives us an idea of how the model performs under every possible threshold value while the value of AUC near 1 would be ideal as we would like to make almost 0% mistakes while identifying the positives. Our group managed to obtain an AUC of 0.8472 (*Figure 3.6)* and the ROC seen in *Figure 3.7***.**

Our team continued to set a threshold of 0.5. If the probability is greater than the threshold of 0.5, it is classified as 1 (churn) ; otherwise, 0 (retained) is given. Following that, our team decided to use the ‘Caret’ package and used the ConfusionMatrix() function to give us the measurements of the confusion matrix of the train set (*Figure*

3.8). Using the train set, the accuracy was given to be 0.795. However, due to our data being unbalanced in positive and negative results, our group deemed accuracy to be an inadequate measurement.

Due to the nature of our project to have a model to predict and attain customers that would actually be retained, our group decided to focus on the measurements on precision, recall (sensitivity) and F1 Score. From *Figure 3.8*, we can see that the precision (positive predictive value) is 0.849 while sensitivity is 0.877. As F1 Score is the harmonic mean of Precision and Recall error metrics for an imbalanced dataset with respect to binary classification of data, our group decided to use the package ‘MLMetrics’ to determine the F1 Score of our models, achieving a F1 score of 0.863. An F1 score is considered to be perfect when it is 1.

Following the above results, our team decided to adjust the threshold levels to optimise our logistic regression model in order to attain a model with the highest f1 score. The table below will show the changes caused by the changes in threshold levels.

Threshold Level	Sensitivity	Precision	F1 Score
0.4	0.802	0.887	0.842
0.5	0.877	0.849	0.863
0.6	0.948	0.796	0.866

From our exploration of threshold levels of 0.4 and 0.6, our team made a few discoveries. By increasing the threshold levels to 0.6, we managed to increase the sensitivity to 0.948 while decreasing the precision to 0.796 (*Figure 3.9*). The F1 score also increased slightly to 0.866 (*Figure 3.9*). Meanwhile, decreasing the threshold level to 0.4 will decrease the sensitivity to 0.802 and increase the precision to 0.887. As for the F1 score, it will decrease to 0.842. These results showed that there was a tradeoff between precision and sensitivity for the different thresholds. Our group has decided that the F1 score would be the determinant for the threshold level chosen due to the weightage it carries for sensitivity and precision, hence have decided that the threshold level of 0.6 was the best threshold value for our logistic regression model.

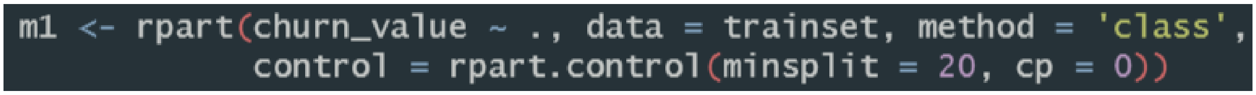
Thereafter, we used the testset and managed to obtain a sensitivity of 0.956 (*Figure 3.11*), precision of 0.792 and F1 Score of 0.866 (*Figure 3.11*).

5.2 Classification and Regression Tree (CART)

CART is a powerful machine learning model that helps to predict categorical outcomes by performing binary splits based on training data. The resulting decision tree is a graphical representation of the probability of how the decision variable is predicted. A key benefit of using decision trees is that they are easy to understand and interpret results. By building a model that can predict the likelihood of customer churn to a high degree of accuracy, we can determine which variables have the biggest impact on causing a customer to leave.

5.2.1 Selection of Model Parameters

For this analysis, the ‘rpart’ and ‘rpart.plot’ packages are used to build the model and display the resulting decision trees respectively. To begin, we use the same train-test split that was earlier used in the logistic regression model to grow the maximal tree. This step is crucial as it allows for fair comparison between the two models. The selected parameters are as follows:



```
m1 <- rpart(churn_value ~ ., data = trainset, method = 'class',
            control = rpart.control(minsplit = 20, cp = 0))
```

Minsplit is a parameter that sets the minimum number of observations in a node before a split can be done. The default value used for minsplits is 20. However, it may not necessarily be optimal, depending on the size of the dataset. While choosing a smaller minsplit increases the resulting number of splits, it also increases the risk of overfitting the CART model to the trainset.

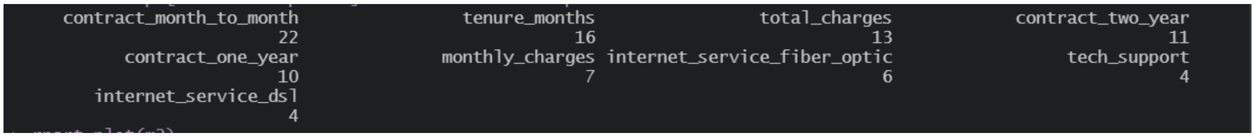
To decide the optimal minsplit, a series of different minsplit values (10, 20, 50, 100) were tested, and the effect on each resulting decision tree was observed. After testing, it was found that selecting a minsplit of 50 and above resulted in a significantly smaller number of splits due to the limited size of the dataset at seven thousand rows. Between 10 and 20, there were no noticeable effects on the resulting node splits and complexity parameters. Thus, it was decided that the optimal minsplit value for this dataset should remain as the default **20**. Since the maximal tree is being grown, the complexity parameter is set at zero.

5.2.2 Finding the Optimal Tree

Next, the maximal tree is pruned to find the optimal tree and reduce overfitting.The optimal tree is a decision tree between the maximal and minimal trees that provides the best generalised model that is neither over nor underfitted. Based on the plot of complexity parameters for the maximal tree (Figure 4.1), it is clear that the second tree is the optimal one. Thus, we prune the maximal tree with the geometric mean of the first two CP values based on the 1 SE rule (0.0215).

5.3.3 Interpretation and Evaluation of Optimal Tree

After pruning, the size of the decision tree has been drastically reduced to only three splits (Figure 4.4). The model has been simplified and is much easier to apply and understand now. The three variables used as splitting points are `contract_month_to_month`, `internet_service_fiber_optic` and `tenure_months`. The splitting criteria for `tenure_months` was >= 16. This suggests that when a customer stays with the telco for more than a year, there is a lower likelihood of customer churn.



```
     contract_month_to_month                tenure_months                total_charges            contract_two_year 
                          22                           16                           13                           11 
           contract_one_year              monthly_charges internet_service_fiber_optic                 tech_support 
                          10                            7                            6                            4 
        internet_service_dsl 
                           4 
```

Variable importance is a measure that shows how much a variable has contributed to the split and is calculated by summing the Gini Index for each split where it is the primary variable and adding the multiple of goodness and adjusted agreement for all splits where it is the surrogate. After scaling the variable importance of the optimal tree out of one hundred, the three most important variables are identified as: **`contract_month_to_month`**, **`tenure_months`** and **`total_charges`**. One interesting observation to note is that while `total_charges` is the third most important variable, it has not been included in the optimal tree. Instead, `internet_service_fiber_optic` was the second split.

Just as with logistic regression, it is necessary to test the predictive accuracy of the CART model on the test split. 'm2' achieved a predictive accuracy of 78.9% and F1 score of 0.867, which is relatively accurate.

5.3 Reasoning for Method of Analysis

To identify the possible reasons for churn, our group has decided to use two prominent statistical methodologies available: logistic regression and classification and regression tree (CART) models. Logistic regression has always been widely known while CART is an emerging tool. Logistic Regression and CART are the simplest yet useful models when it comes to predictive analysis. However, they differ in their own special ways.

From our research, we have found that logistic regression, despite being widely popular and easy to use, seems to have a large challenge in which is the difficulty to interpret the results correctly and precisely. For example, from LR2 in logistic regression, it is very difficult to understand what is the cause of churn. Meanwhile, from m2, we can almost immediately identify the factors that affect churn, with **`contract_month_to_month`** being the primary factor.

Next, due to the above limitation of logistic regression, it could lead to the wrong variables being picked for logistic regression, leading to a domino effect of a wrong analysis, showing that CART is a much safer choice. However, this does not undermine the fact that CART has its own limitations such as the cost of the sample size. Every time it splits the data using a predictor, the sample size reduces significantly as compared to logistic regression where it is looking at the simultaneous effects of all predictors.

For our group, due to the extensive size of our dataset, our group deemed the effects' of CART's limitation to be minimal. In addition, from the F1_Score of both CART and Logistic Regression, we could observe that the CART model is much more accurate as compared to logistic regression.

	F1_Score	Accuracy%
Logistic Regression	0.866	78.4%
CART	0.867	78.9

To sum up, our group has chosen to use CART as our primary predictive model, but would not be limited to having logistic regression as our alternative model.

6. Recommended Solutions and Implementations

6.1 Customer Loyalty Program

As mentioned in Section 4.4 above, we noticed many customers tend to opt for month-to-month contracts instead as **`Contract_month_to_month`** seems to be the most statistically significant variable. This could be attributed to the fact that most customers nowadays dislike being tied down by a particular contract and enjoy the freedom of not being obligated to the terms of their contract (Llopis, 2014). This could also be a differentiating factor compared to the other telcos, such that customers are free to terminate the contract immediately if they are dissatisfied with the services.

Although M1 does have a minimum commitment period of 1 month only, we noticed there are many hidden costs such as early re-contract fee and re-contract administrative fee. These costs might deter customers from signing contracts with M1. For example, the early re-contract fee ranges from $200 to $350 if customers have yet to fulfil more than 21 months of their contract. With reputable phone companies like Apple coming out with new models within a span of a few months, it is not surprising for customers to want to keep up with the newest models. Perhaps M1 can also consider lowering the fees and come up with new plans constantly in order to entice customers to sign contracts with them.

Currently, M1 has a Sunrisers Programme which is a rewards programme that offers exclusive privileges to their valued customers. However, this programme does not entice many people to sign up due to its preposterous requirements such as average spend of $300 over the past 6 months and the membership is strictly by invitation.

In order to reduce the customer churn for M1, our group would like to propose for M1 to incorporate the following customer loyalty program based on our analysis.

1. Offers & Rewards
    a. Variety of Rewards

The ultimate formula for people to be enticed into a loyalty program would be to have a wide array of rewards that are attractive to the customers' passions. M1 would have to provide more than just free additional mobile data or caller id to customers. For example, M1 could offer a free staycation or a pair of tickets to Singapore's renowned attractions or even a limited-time subscription to streaming platforms. These offers and rewards could be incorporated for members who have opted to sign with M1 for longer contracts.

b. Surprise

M1 could utilise their mobile application and give their customers an unanticipated surprise by giving them a reward for their birthday or membership anniversary. This will help give M1 a tinge of personal touch, and hence allow the customers to feel more valued.

c. Tiers

M1 could have an enhanced tiered program for their customers. For instance, for every dollar spent on M1, they could unlock different tiers, thus different rewards. By having a tiered program, M1 will be able to make people feel rewarded for their loyalty, hence ultimately reducing customer churn. Moreover, this could induce customers on **`Contract_Month_to_Month`** to continue staying with M1 after every month to climb the respective tiers.

2. Personalisation

a. Customer Profiling

M1 could incorporate rewards in surveys and tasks in their applications to entice their customers to participate in a loyalty program. This would be beneficial to M1 in many ways as incentivised customer profiling would allow M1 to learn more about the customers' characteristics, values, lifestyle, behaviours and rewards they like, passion etc, allowing for M1 to introduce personal touch to their customers.

6.2 Bundling

Adding onto our loyalty program, our group believes that M1 could enhance their bundles offered to the customers.

Currently, M1 offers direct carrier billing in which customers are allowed to charge their subscriptions or purchases directly to their monthly bills. However, the price of digital services provided by M1 remains the same as what the service providers normally offer. For instance, customers who subscribed to Netflix through M1 pay the same amount as those who are not a customer of M1. Such a carrier billing system is not sufficient to differentiate M1 from the other Telco in Singapore which further explains why some customers might choose to churn and switch to another Telco which offers a lower price for the same services. Therefore, M1 should partner with, say some entertainment streaming providers, to give their customers exclusive promotional offers on these streaming services. For example, M1 could offer a complimentary Netflix subscription if the customer signs up for a certain package under M1.

Moreover, M1 could create a “triple play” or even a “quadruple play” of bundles in which multiple services are bundled together and offered to customers with incentives. For instance, by amending their traditional bundle offerings of talk time and data to include fibre broadband service with some TV or movies streaming services, such as Disney+ and Netflix. The two crucial factors that drive customers' preference for bundles are convenience and cost savings. With such offerings of bundled services, M1 customers are able to pay for multiple services under a single bill. They also get to enjoy greater savings under such a multiplay package than to purchase each service separately. Since bundling creates switching costs, customers are therefore less likely to churn which in turn enhances customer loyalty with M1.

Our team also further considered the fact that it is not financially viable for most customers to subscribe to that many services. Therefore, M1 could customize multiple bundle options for every segment of its customer base. For instance, M1 could offer a Sports Plus pack for millennials who are enthusiastic about sports and like to watch live sports regularly. What M1 could also possibly do is to adjust the prices of different bundles according to their customer segments reflecting the difference in valuations for their commodities. For example, M1 could offer discounts on data communication fees for students to mitigate the extra charges for students taking online classes during the current pandemic period.

7. Conclusion

As competition increases and the telecommunications industry continues to evolve, telcos should focus more of their resources on reducing customer churn. Telecommunication capabilities are essential for living in the internet age, and customers who leave M1 will more than likely take their business to a competitor, be it MVNO or MNO. The results of our analysis show that while customer churn is a complex issue, there are certain variables that have a higher influence on the customer's likelihood to churn. Therefore, it is critical for telcos to prioritise on retaining customers by understanding the key factors affecting customer churn.

To identify these factors, we applied machine learning techniques to a dataset from a telco based in California, USA. Understandably, there are certain limitations as there was a critical assumption that customers in the USA would have roughly the same consumer behaviour as those in Singapore.

A two-pronged machine learning approach was taken to analyze the problem. Firstly, logistic regression was used on the dataset to develop a predictive model with the highest F1 score, which is a harmonic mean of the precision and recall. Additionally, odds ratio was used as a measure of variable importance. Secondly, a CART model was built to identify the most important variables involved in splitting the optimal decision tree. From our findings, both methods have shown that the key variable affecting customer churn decision is **`contract_month_to_month`**.

After our analysis, we came up with solutions to tackle the problem of customer churn in telco companies. Firstly, the use of customer loyalty programs will help elevate the customer's experience. Customised offers and rewards catered specifically to their needs can also be used to retain customers in the long run. In addition, the use of bundling services together can be used to entice customers to engage in M1's services since they will be able to save costs as well.

All in all, it is not possible to find a one-size-fits-all solution to the problems faced by telcos companies like M1. However, with the help of machine learning models, we are able to delve deeper into the root cause of our identified problem and come up with various solutions to target it.

Appendices

Annex A: Data Cleaning

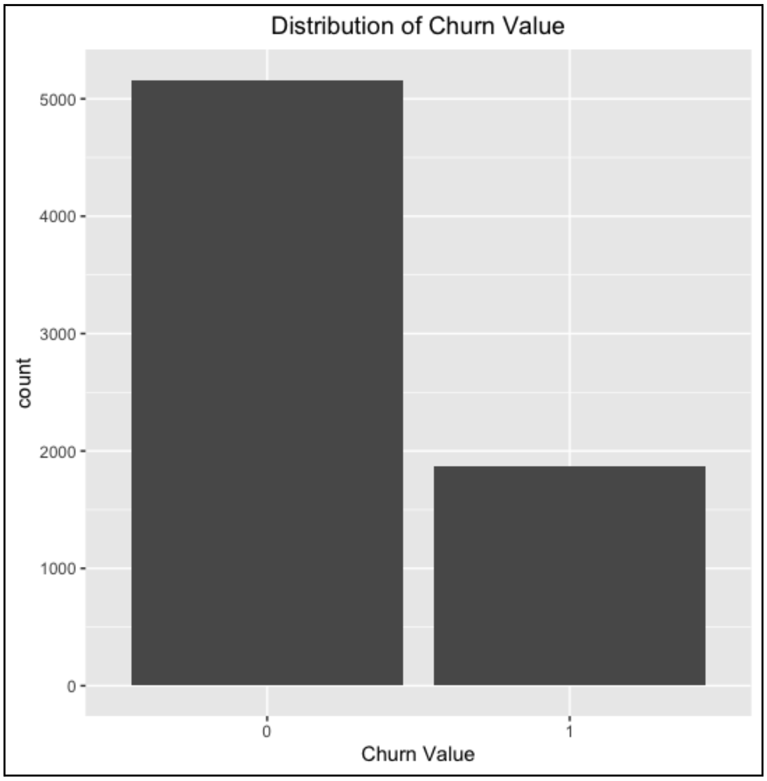


Figure 1: Imbalance in decision variable

Annex B: Data Exploration & Visualization

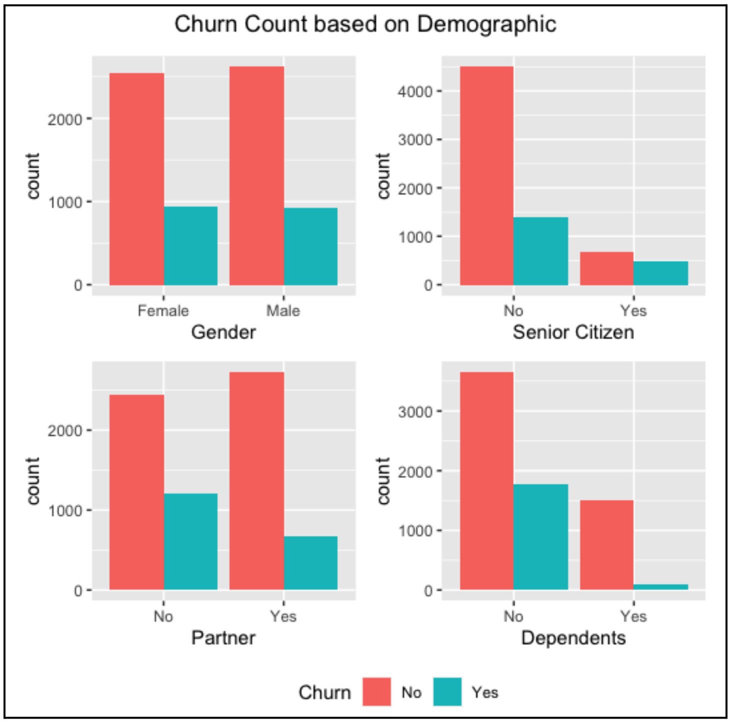


Figure 2.1: Demographic Categorical Variables

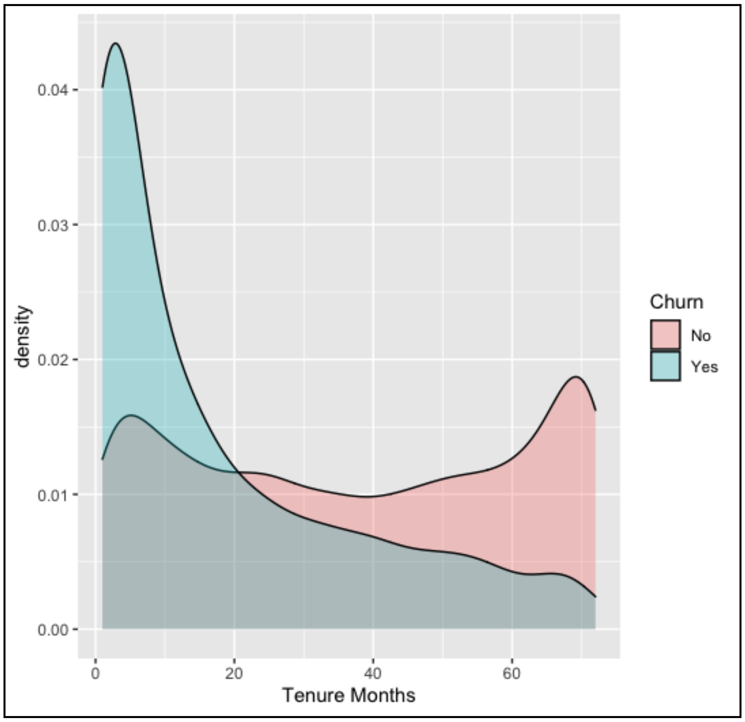


Figure 2.2: Distribution of Tenure

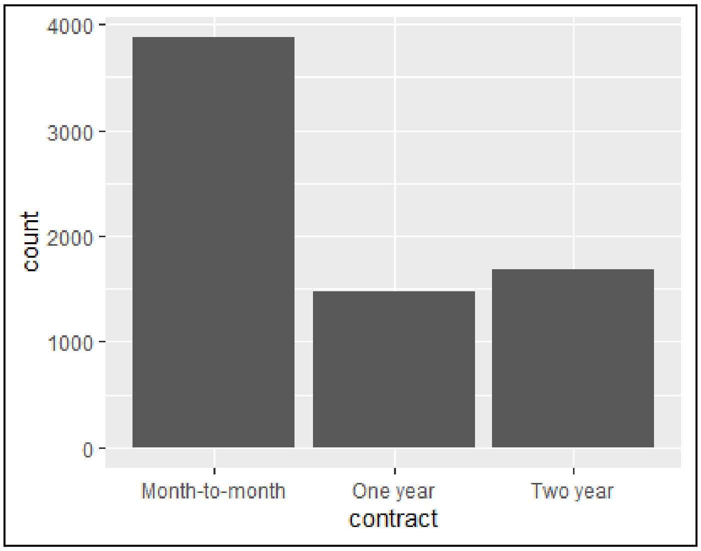


Figure 2.3: Proportions of customers with different types of contract

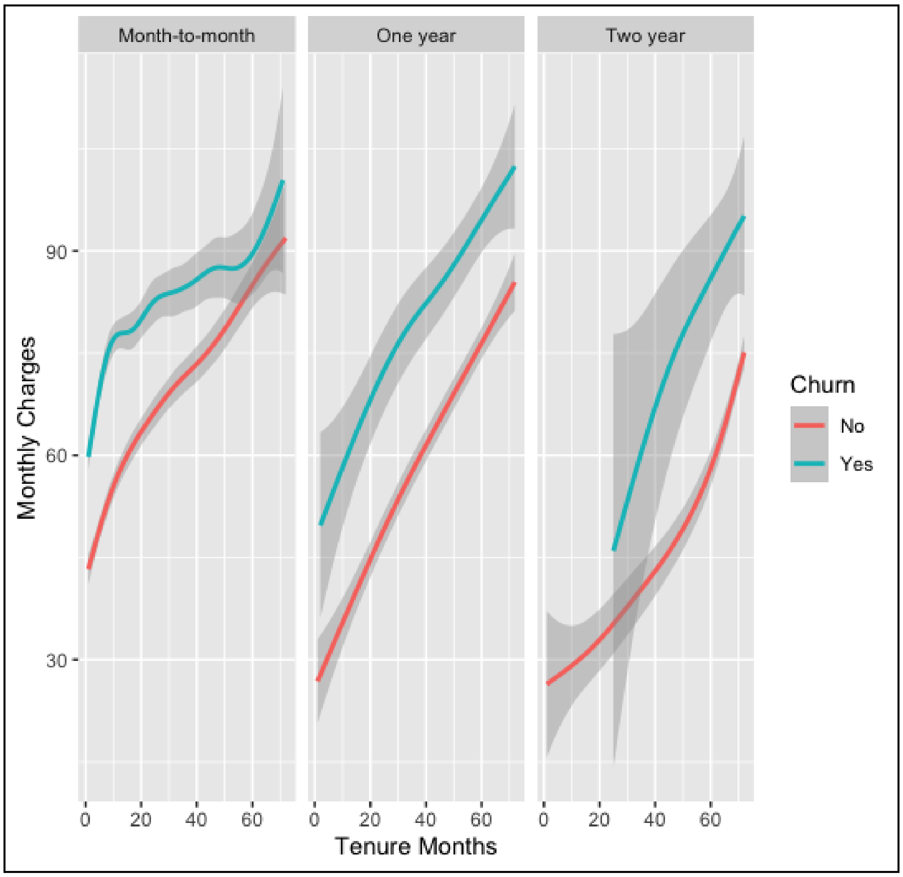


Figure 2.4: Monthly Charges across Contracts

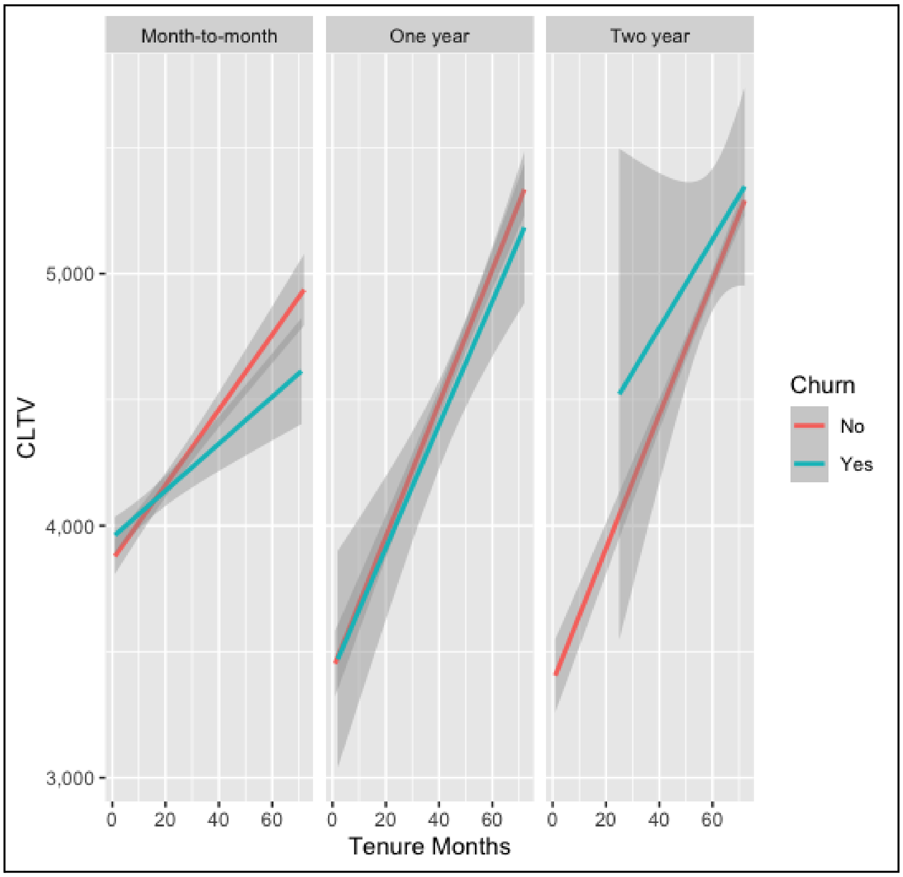


Figure 2.5: Customer Lifetime Value across Contracts

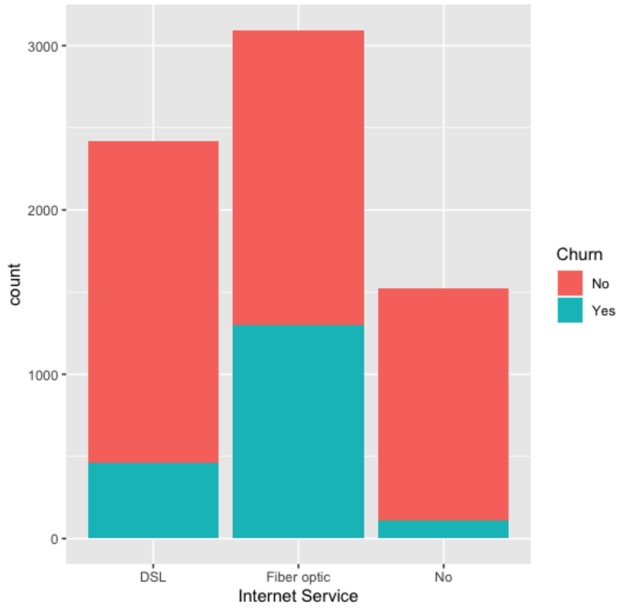


Figure 2.6(a): Internet Service Subscription

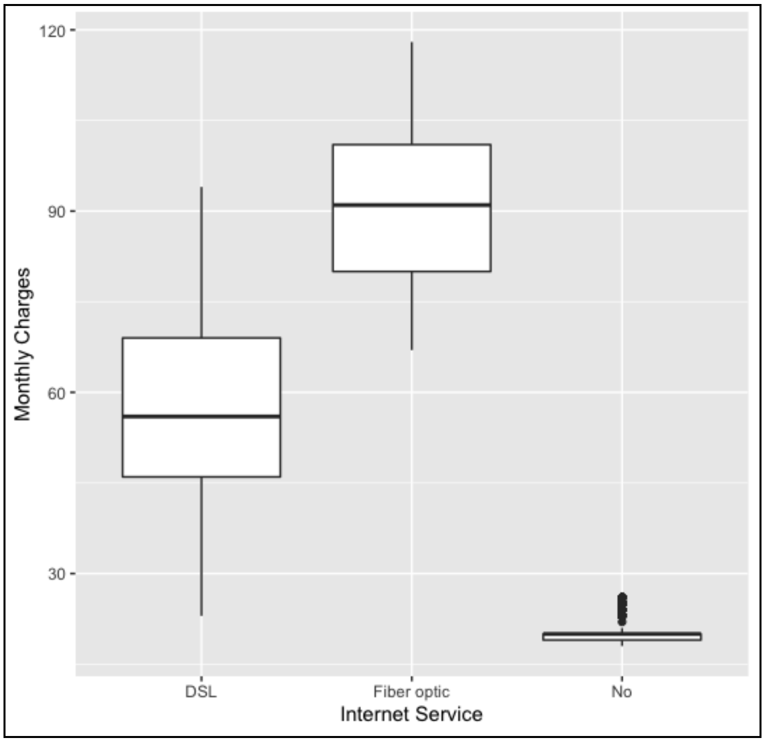


Figure 2.6(b): Monthly Charges across Internet Services

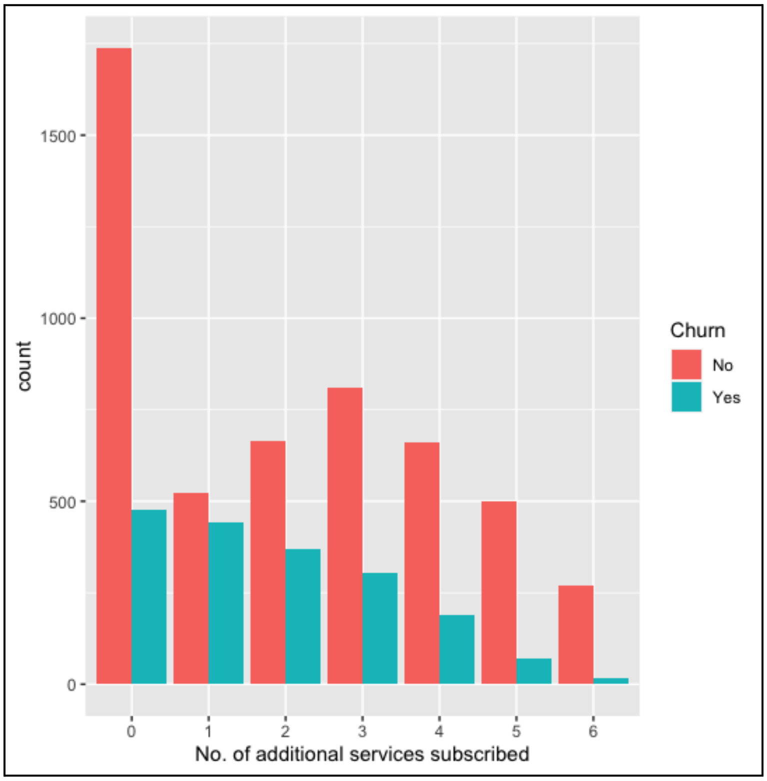
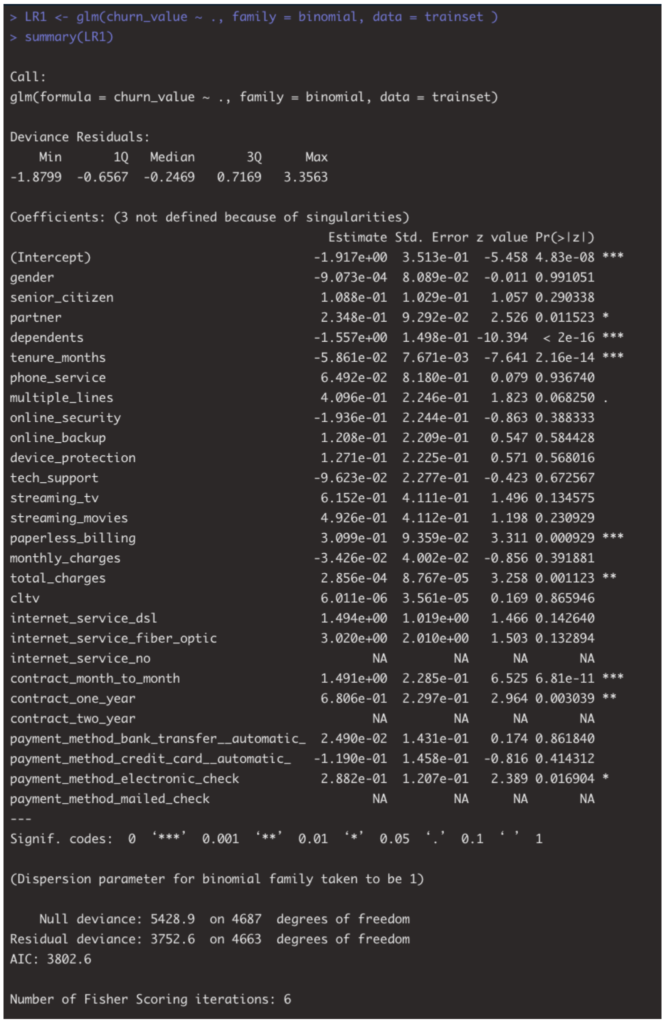


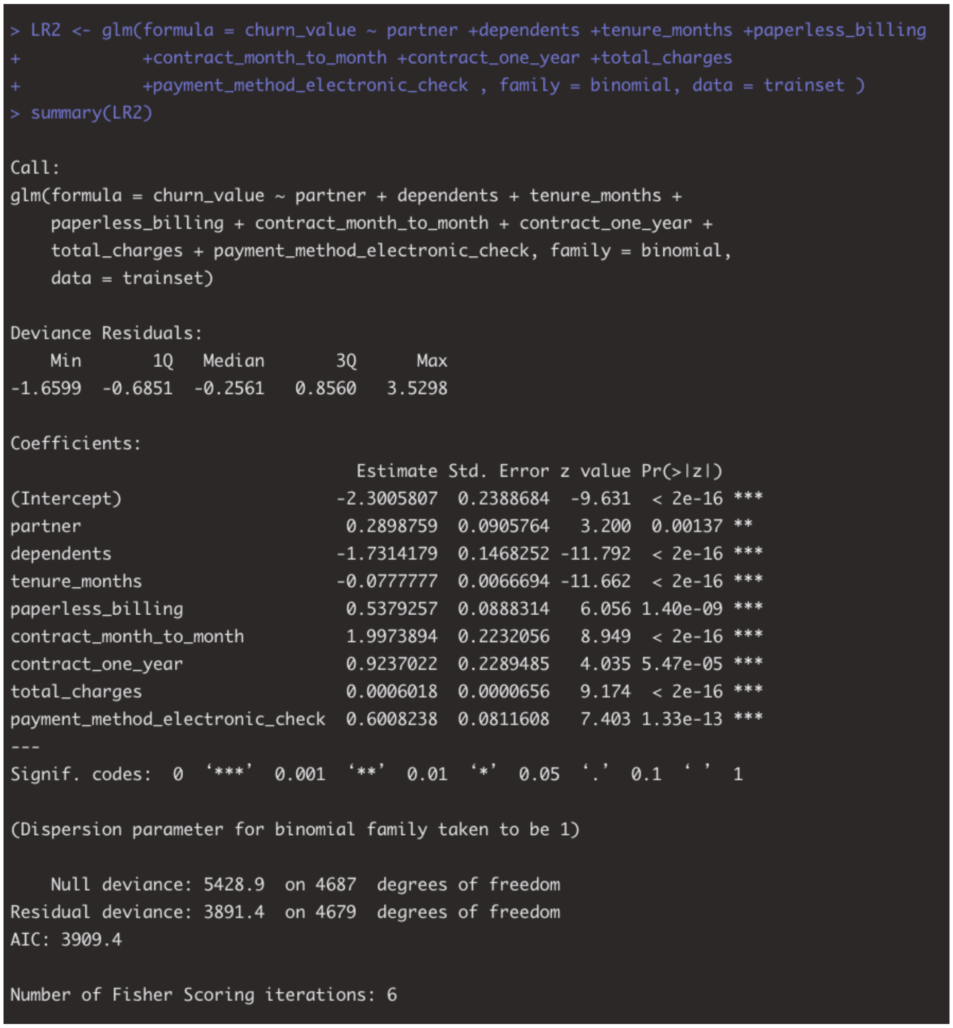
Figure 2.7: Additional Services

Annex C: Logistic Regression



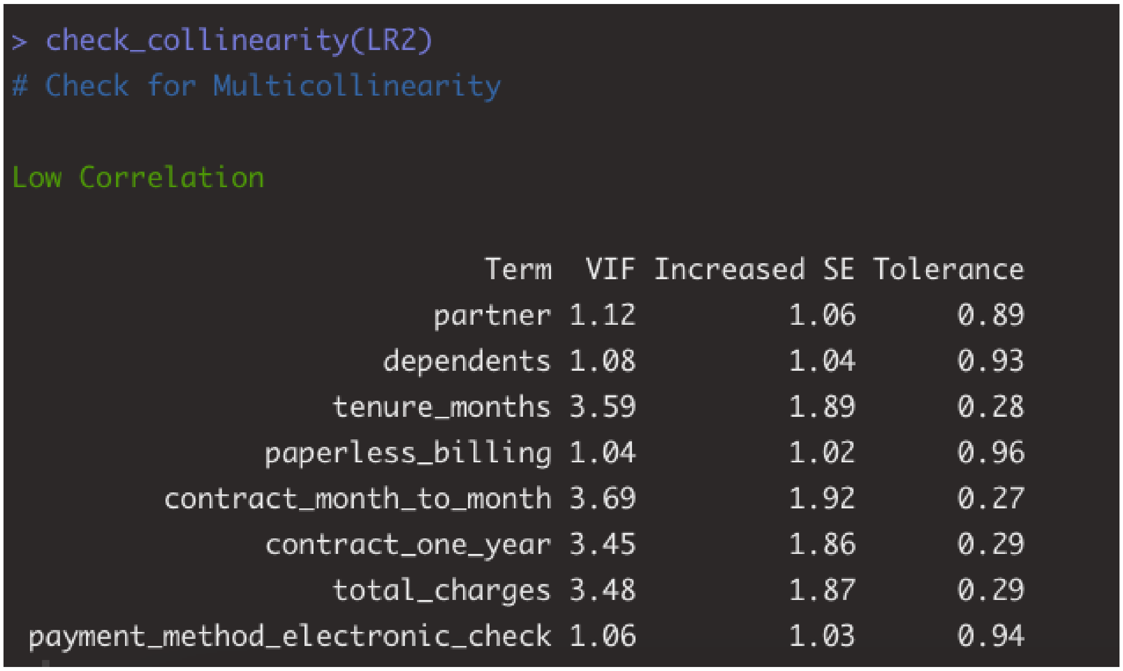
```
> LR1 <- glm(churn_value ~ ., family = binomial, data = trainset )
> summary(LR1)

Call:
glm(formula = churn_value ~ ., family = binomial, data = trainset)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.8799  -0.6567  -0.2469   0.7169   3.3563  

Coefficients: (3 not defined because of singularities)
                                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)                              -1.917e+00  3.513e-01  -5.458 4.83e-08 ***
gender                                   -9.073e-04  8.089e-02  -0.011 0.991051    
senior_citizen                            1.088e-01  1.029e-01   1.057 0.290338    
partner                                   2.348e-01  9.292e-02   2.526 0.011523 *  
dependents                               -1.557e+00  1.498e-01 -10.394  < 2e-16 ***
tenure_months                            -5.861e-02  7.671e-03  -7.641 2.16e-14 ***
phone_service                             6.492e-02  8.180e-01   0.079 0.936740    
multiple_lines                            4.096e-01  2.246e-01   1.823 0.068250 .  
online_security                          -1.936e-01  2.244e-01  -0.863 0.388333    
online_backup                             1.208e-01  2.209e-01   0.547 0.584428    
device_protection                         1.271e-01  2.225e-01   0.571 0.568016    
tech_support                             -9.623e-02  2.277e-01  -0.423 0.672567    
streaming_tv                              6.152e-01  4.111e-01   1.496 0.134575    
streaming_movies                          4.926e-01  4.112e-01   1.198 0.230929    
paperless_billing                         3.099e-01  9.359e-02   3.311 0.000929 ***
monthly_charges                          -3.426e-02  4.002e-02  -0.856 0.391881    
total_charges                             2.856e-04  8.767e-05   3.258 0.001123 ** 
cltv                                      6.011e-06  3.561e-05   0.169 0.865946    
internet_service_dsl                      1.494e+00  1.019e+00   1.466 0.142640    
internet_service_fiber_optic              3.020e+00  2.010e+00   1.503 0.132894    
internet_service_no                              NA         NA      NA       NA    
contract_month_to_month                   1.491e+00  2.285e-01   6.525 6.81e-11 ***
contract_one_year                         6.806e-01  2.297e-01   2.964 0.003039 ** 
contract_two_year                                NA         NA      NA       NA    
payment_method_bank_transfer__automatic_  2.490e-02  1.431e-01   0.174 0.861840    
payment_method_credit_card__automatic_   -1.190e-01  1.458e-01  -0.816 0.414312    
payment_method_electronic_check           2.882e-01  1.207e-01   2.389 0.016904 *  
payment_method_mailed_check                      NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5428.9  on 4687  degrees of freedom
Residual deviance: 3752.6  on 4663  degrees of freedom
AIC: 3802.6

Number of Fisher Scoring iterations: 6
```

Figure 3.1: LR1



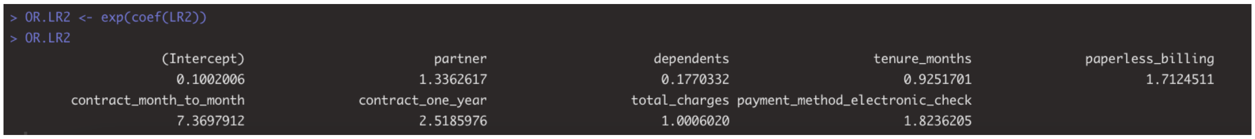
```
> LR2 <- glm(formula = churn_value ~ partner +dependents +tenure_months +paperless_billing
+             +contract_month_to_month +contract_one_year +total_charges
+             +payment_method_electronic_check , family = binomial, data = trainset )
> summary(LR2)

Call:
glm(formula = churn_value ~ partner + dependents + tenure_months +
    paperless_billing + contract_month_to_month + contract_one_year +
    total_charges + payment_method_electronic_check, family = binomial,
    data = trainset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6599  -0.6851  -0.2561   0.8560   3.5298

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -2.3005807  0.2388684  -9.631  < 2e-16 ***
partner                          0.2898759  0.0905764   3.200  0.00137 **
dependents                      -1.7314179  0.1468252 -11.792  < 2e-16 ***
tenure_months                   -0.0777777  0.0066694 -11.662  < 2e-16 ***
paperless_billing                0.5379257  0.0888314   6.056 1.40e-09 ***
contract_month_to_month          1.9973894  0.2232056   8.949  < 2e-16 ***
contract_one_year                0.9237022  0.2289485   4.035 5.47e-05 ***
total_charges                    0.0006018  0.0000656   9.174  < 2e-16 ***
payment_method_electronic_check  0.6008238  0.0811608   7.403 1.33e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5428.9  on 4687  degrees of freedom
Residual deviance: 3891.4  on 4679  degrees of freedom
AIC: 3909.4

Number of Fisher Scoring iterations: 6
```

Figure 3.2: LR2



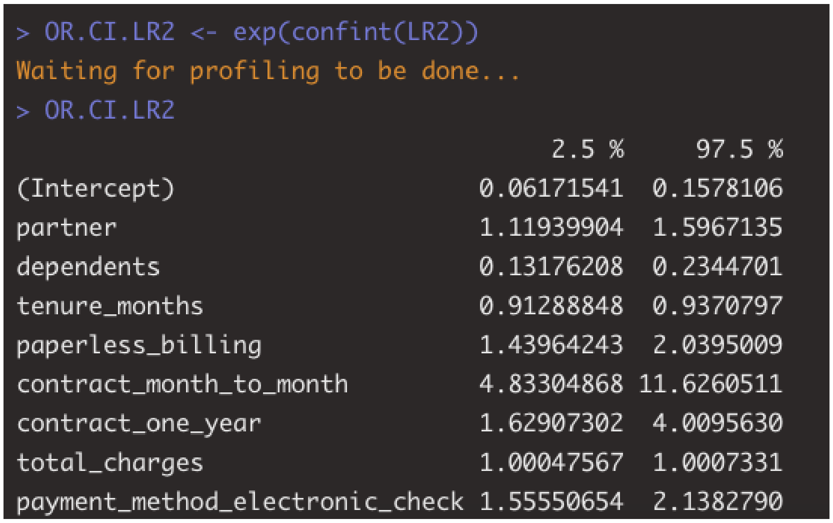
```
> check_collinearity(LR2)
# Check for Multicollinearity

Low Correlation

                            Term  VIF Increased SE Tolerance
                         partner 1.12         1.06      0.89
                      dependents 1.08         1.04      0.93
                   tenure_months 3.59         1.89      0.28
               paperless_billing 1.04         1.02      0.96
         contract_month_to_month 3.69         1.92      0.27
               contract_one_year 3.45         1.86      0.29
                   total_charges 3.48         1.87      0.29
 payment_method_electronic_check 1.06         1.03      0.94
```

Figure 3.3: Multicollinearity



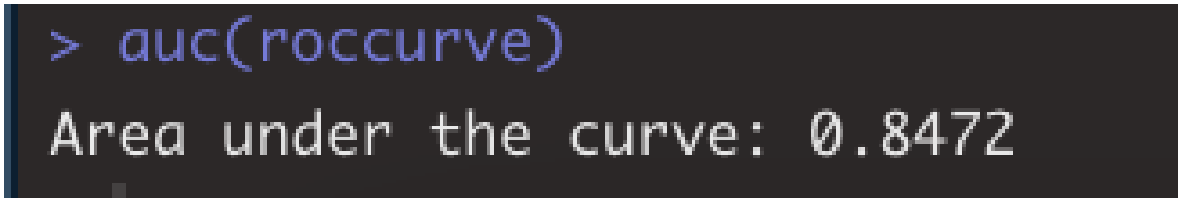
```
> OR.LR2 <- exp(coef(LR2))
> OR.LR2
                    (Intercept)                         partner                      dependents                   tenure_months               paperless_billing
                      0.1002006                       1.3362617                       0.1770332                       0.9251701                       1.7124511
        contract_month_to_month               contract_one_year                   total_charges payment_method_electronic_check
                      7.3697912                       2.5185976                       1.0006020                       1.8236205
```

Figure 3.4: Odds Ratio



```
> OR.CI.LR2 <- exp(confint(LR2))
Waiting for profiling to be done...
> OR.CI.LR2
                                     2.5 %     97.5 %
(Intercept)                     0.06171541  0.1578106
partner                         1.11939904  1.5967135
dependents                      0.13176208  0.2344701
tenure_months                   0.91288848  0.9370797
paperless_billing               1.43964243  2.0395009
contract_month_to_month         4.83304868 11.6260511
contract_one_year               1.62907302  4.0095630
total_charges                   1.00047567  1.0007331
payment_method_electronic_check 1.55550654  2.1382790
```

Figure 3.5: Odds Ratio Confidence Interval Level



```
> auc(roccurve)
Area under the curve: 0.8472
```

Figure 3.6: AUC

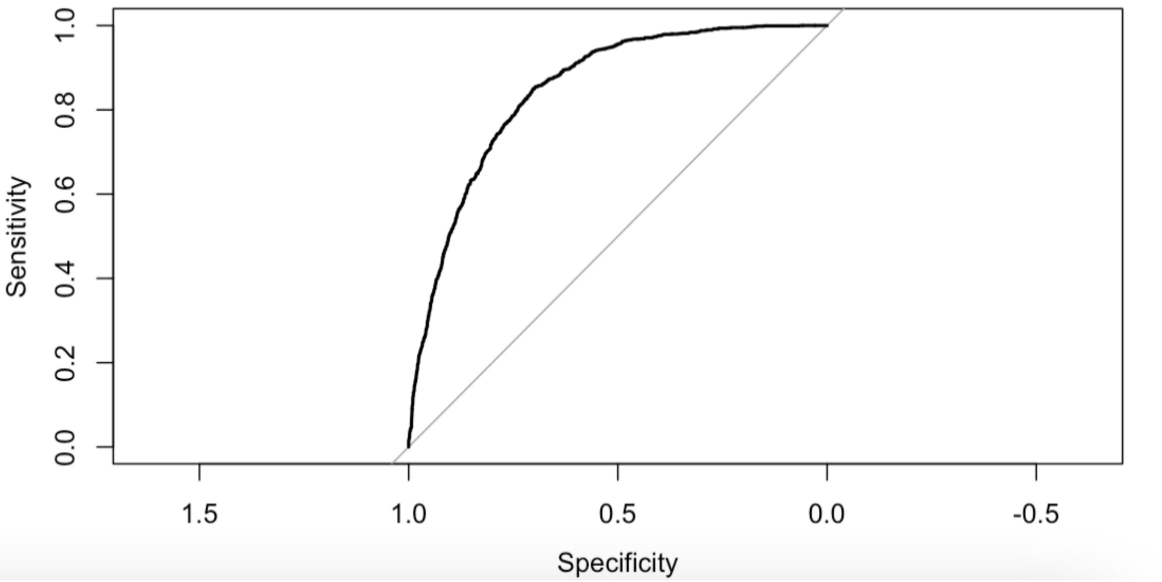
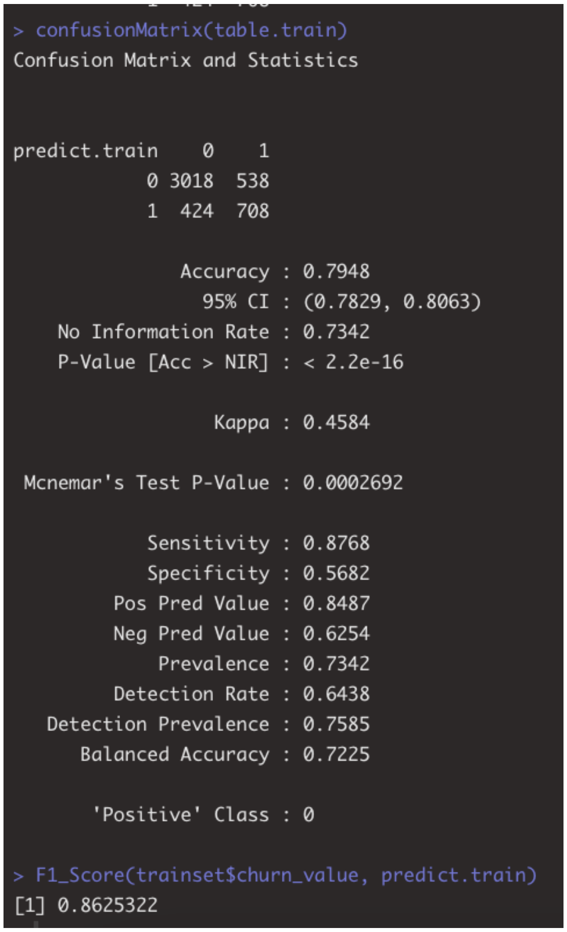
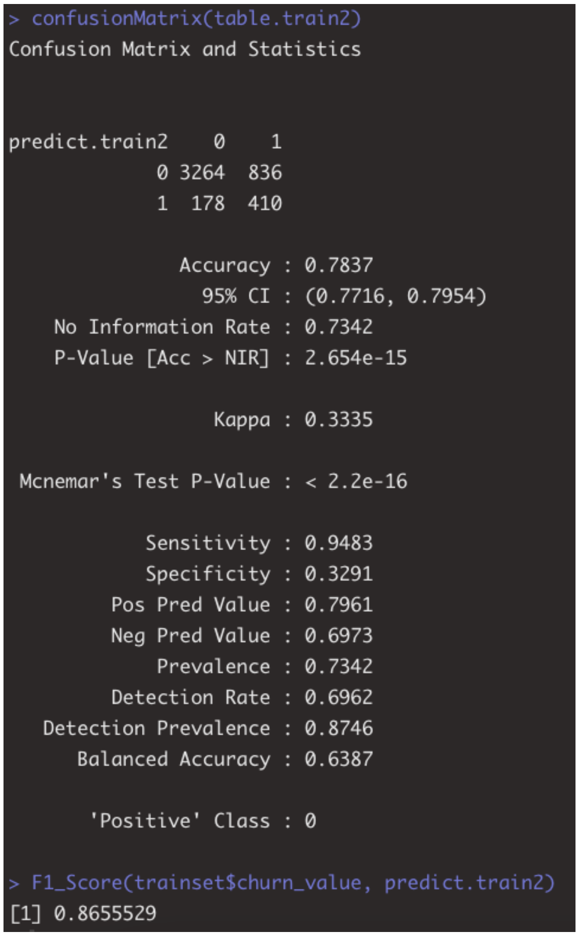


Figure 3.7: ROC



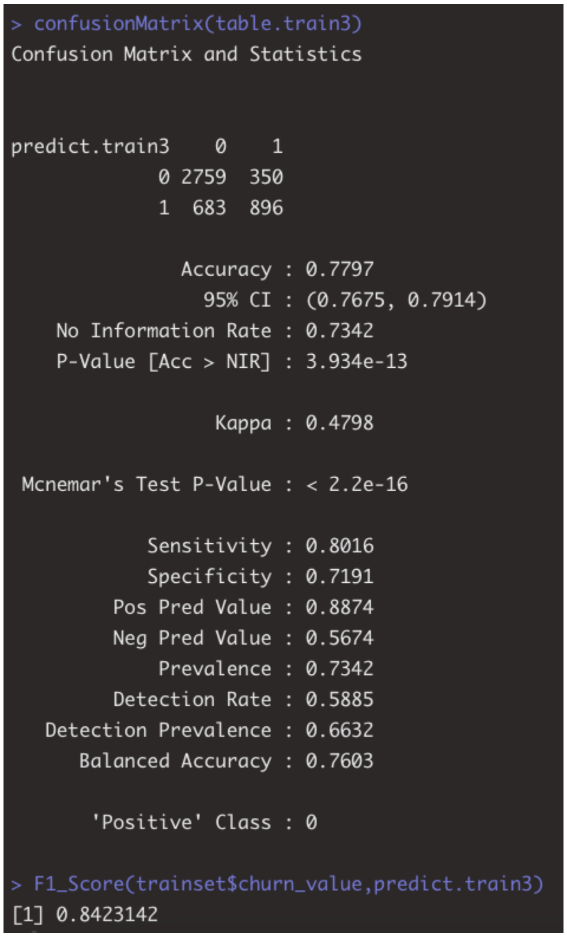
```
> confusionMatrix(table.train)
Confusion Matrix and Statistics


predict.train    0    1
            0 3018  538
            1  424  708

               Accuracy : 0.7948
                 95% CI : (0.7829, 0.8063)
    No Information Rate : 0.7342
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4584

 Mcnemar's Test P-Value : 0.0002692

            Sensitivity : 0.8768
            Specificity : 0.5682
         Pos Pred Value : 0.8487
         Neg Pred Value : 0.6254
             Prevalence : 0.7342
         Detection Rate : 0.6438
   Detection Prevalence : 0.7585
      Balanced Accuracy : 0.7225

       'Positive' Class : 0

> F1_Score(trainset$churn_value, predict.train)
[1] 0.8625322
```

Figure 3.8: Confusion Matrix and F1 score of table.train



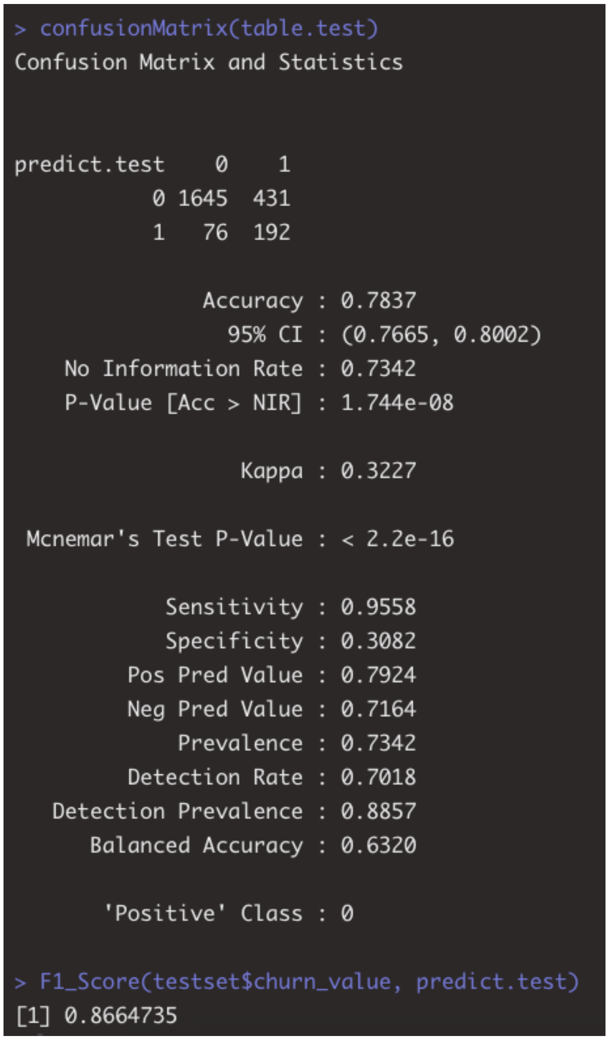
```
> confusionMatrix(table.train2)
Confusion Matrix and Statistics


predict.train2    0    1
             0 3264  836
             1  178  410

               Accuracy : 0.7837
                 95% CI : (0.7716, 0.7954)
    No Information Rate : 0.7342
    P-Value [Acc > NIR] : 2.654e-15

                  Kappa : 0.3335

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9483
            Specificity : 0.3291
         Pos Pred Value : 0.7961
         Neg Pred Value : 0.6973
             Prevalence : 0.7342
         Detection Rate : 0.6962
   Detection Prevalence : 0.8746
      Balanced Accuracy : 0.6387

       'Positive' Class : 0

> F1_Score(trainset$churn_value, predict.train2)
[1] 0.8655529
```

Figure 3.9: Confusion Matrix of and F1 score of table.train2



```
> confusionMatrix(table.train3)
Confusion Matrix and Statistics


predict.train3    0    1
             0 2759  350
             1  683  896

               Accuracy : 0.7797
                 95% CI : (0.7675, 0.7914)
    No Information Rate : 0.7342
    P-Value [Acc > NIR] : 3.934e-13

                  Kappa : 0.4798

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8016
            Specificity : 0.7191
         Pos Pred Value : 0.8874
         Neg Pred Value : 0.5674
             Prevalence : 0.7342
         Detection Rate : 0.5885
   Detection Prevalence : 0.6632
      Balanced Accuracy : 0.7603

       'Positive' Class : 0

> F1_Score(trainset$churn_value,predict.train3)
[1] 0.8423142
```

Figure 3.10: Confusion matrix of and F1 score of table.train3



```
> confusionMatrix(table.test)
Confusion Matrix and Statistics


predict.test    0    1
           0 1645  431
           1   76  192

               Accuracy : 0.7837
                 95% CI : (0.7665, 0.8002)
    No Information Rate : 0.7342
    P-Value [Acc > NIR] : 1.744e-08

                  Kappa : 0.3227

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9558
            Specificity : 0.3082
         Pos Pred Value : 0.7924
         Neg Pred Value : 0.7164
             Prevalence : 0.7342
         Detection Rate : 0.7018
   Detection Prevalence : 0.8857
      Balanced Accuracy : 0.6320

       'Positive' Class : 0

> F1_Score(testset$churn_value, predict.test)
[1] 0.8664735
```

Figure 3.11 : Confusion matrix and F1 score of table.test

Annex D: CART

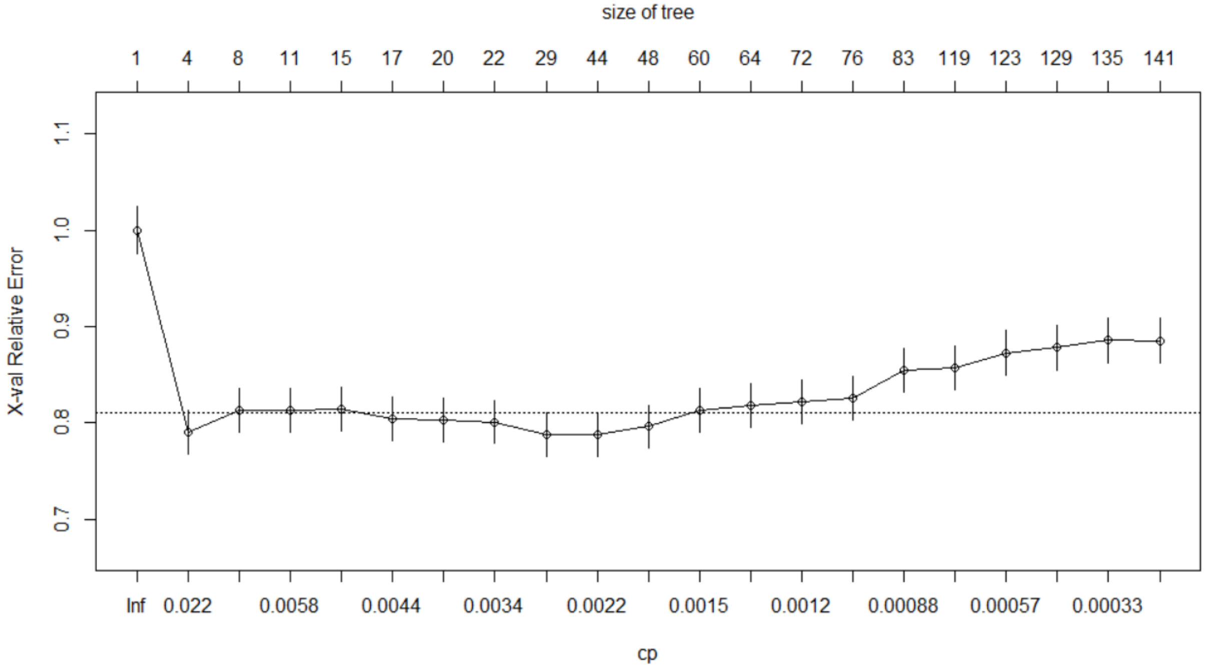


Figure 4.1: Complexity Parameter Plot of Maximal Tree

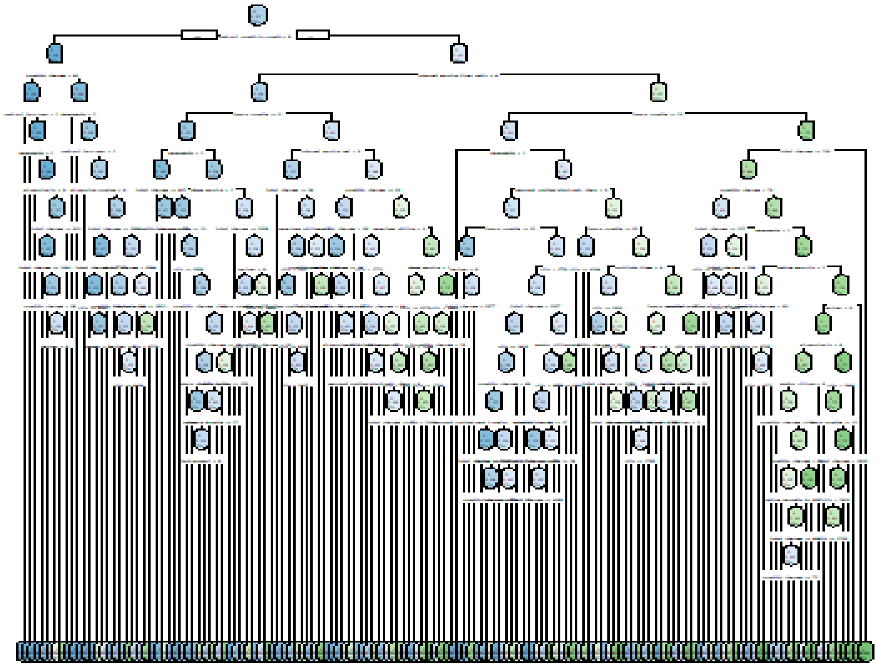
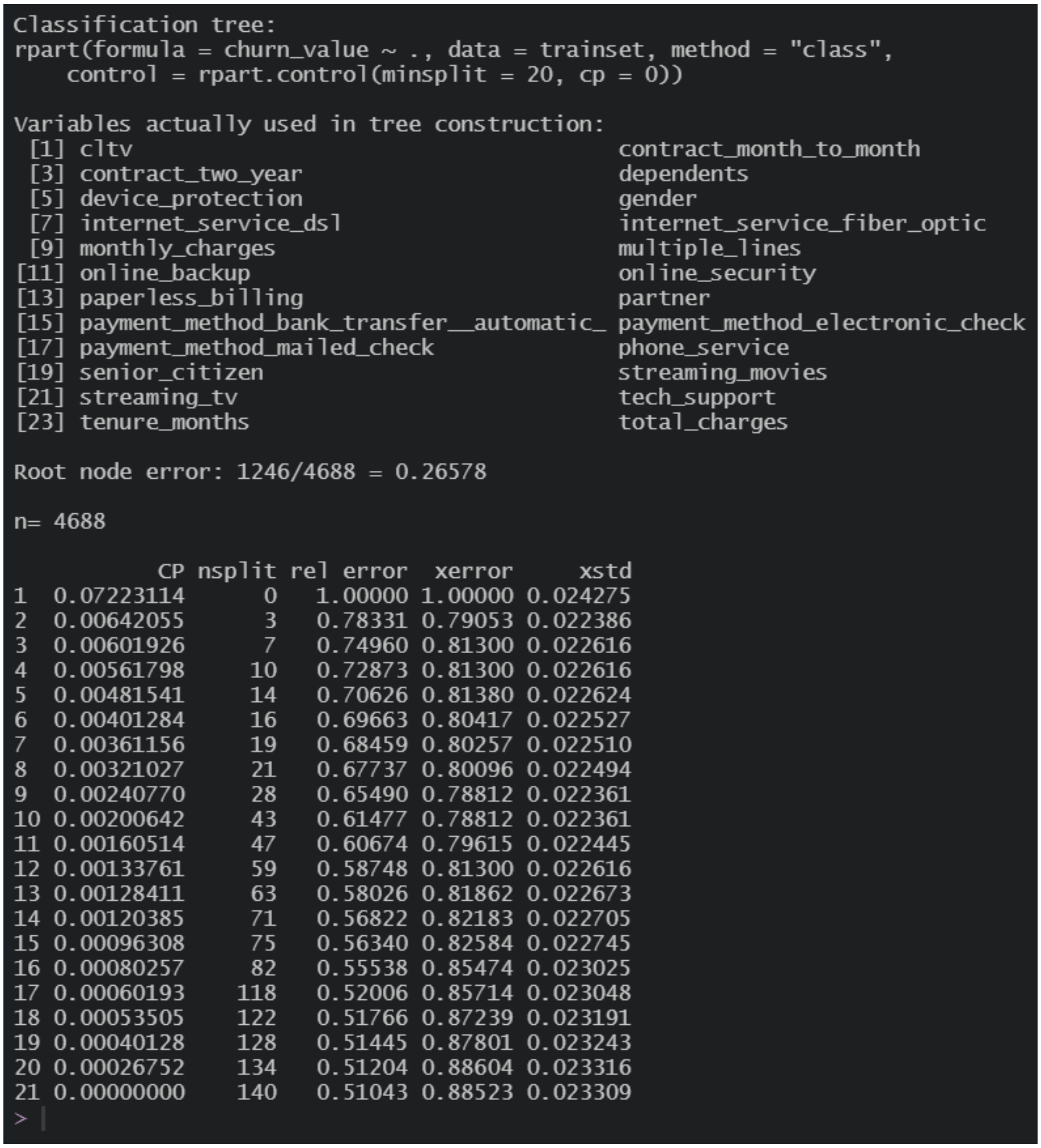


Figure 4.2: Maximal Tree 'm1'



```
Classification tree:
rpart(formula = churn_value ~ ., data = trainset, method = "class",
    control = rpart.control(minsplit = 20, cp = 0))

Variables actually used in tree construction:
 [1] cltv                                  contract_month_to_month
 [3] contract_two_year                     dependents
 [5] device_protection                     gender
 [7] internet_service_dsl                  internet_service_fiber_optic
 [9] monthly_charges                       multiple_lines
[11] online_backup                         online_security
[13] paperless_billing                     partner
[15] payment_method_bank_transfer__automatic_ payment_method_electronic_check
[17] payment_method_mailed_check           phone_service
[19] senior_citizen                        streaming_movies
[21] streaming_tv                          tech_support
[23] tenure_months                         total_charges

Root node error: 1246/4688 = 0.26578

n= 4688

           CP nsplit rel error  xerror     xstd
1  0.07223114      0   1.00000 1.00000 0.024275
2  0.00642055      3   0.78331 0.79053 0.022386
3  0.00601926      7   0.74960 0.81300 0.022616
4  0.00561798     10   0.72873 0.81300 0.022616
5  0.00481541     14   0.70626 0.81380 0.022624
6  0.00401284     16   0.69663 0.80417 0.022527
7  0.00361156     19   0.68459 0.80257 0.022510
8  0.00321027     21   0.67737 0.80096 0.022494
9  0.00240770     28   0.65490 0.78812 0.022361
10 0.00200642     43   0.61477 0.78812 0.022361
11 0.00160514     47   0.60674 0.79615 0.022445
12 0.00133761     59   0.58748 0.81300 0.022616
13 0.00128411     63   0.58026 0.81862 0.022673
14 0.00120385     71   0.56822 0.82183 0.022705
15 0.00096308     75   0.56340 0.82584 0.022745
16 0.00080257     82   0.55538 0.85474 0.023025
17 0.00060193    118   0.52006 0.85714 0.023048
18 0.00053505    122   0.51766 0.87239 0.023191
19 0.00040128    128   0.51445 0.87801 0.023243
20 0.00026752    134   0.51204 0.88604 0.023316
21 0.00000000    140   0.51043 0.88523 0.023309
> |
```

Figure 4.3: Complexity Parameter and 10-CV Error for ‘m1’

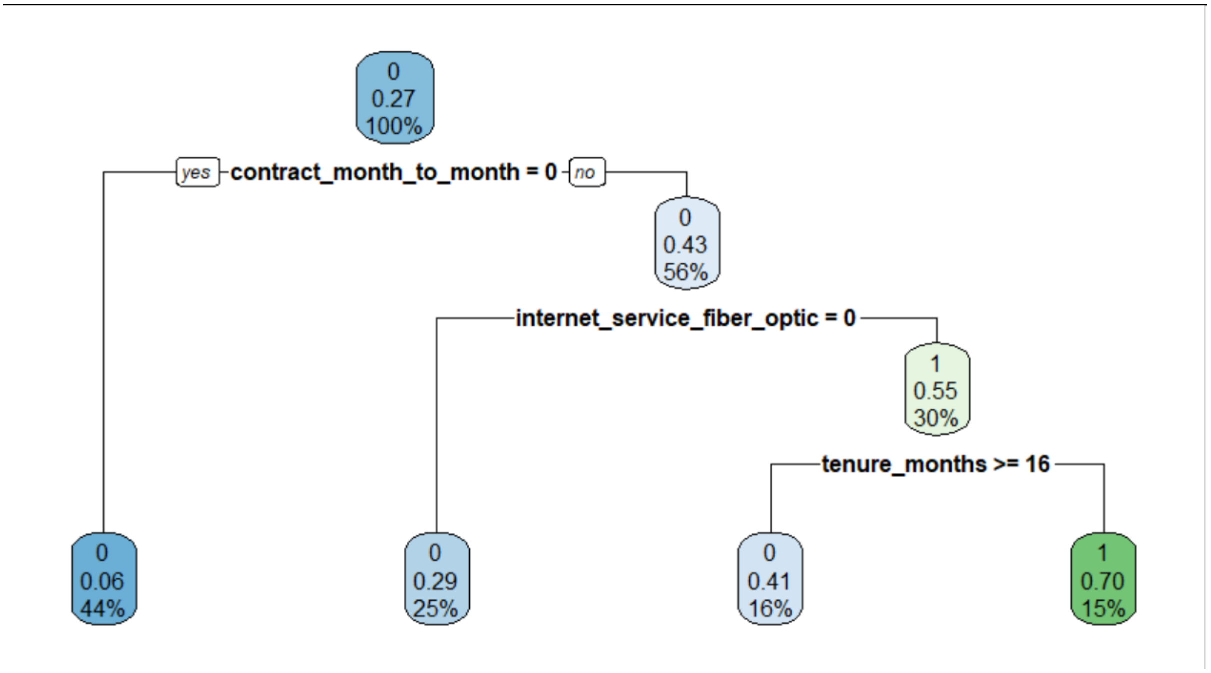
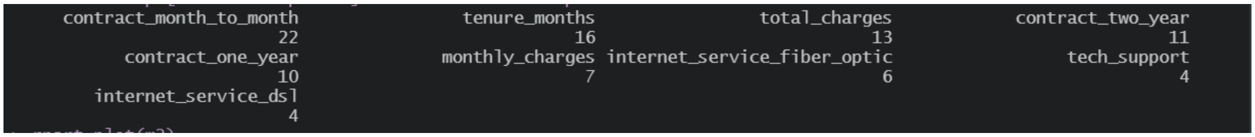


Figure 4.4: Optimal Tree ‘m2’



```
  contract_month_to_month               tenure_months               total_charges           contract_two_year
                       22                          16                          13                          11
        contract_one_year             monthly_charges internet_service_fiber_optic                tech_support
                       10                           7                           6                           4
     internet_service_dsl
                        4
```

Figure 4.5: Scaled Variable Importance for ‘m2’

References

Data source:
https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113

Reichheld, F. and Sasser, Jr., W., 1990. *Zero Defections: Quality Comes to Services*. [online] Harvard Business Review. Available at: <https://hbr.org/1990/09/zero-defections-quality-comes-to-services>

Brownlee, J. (2020, January 14). *A gentle introduction to imbalanced classification*. Machine Learning Mastery. Retrieved November 3, 2021, from https://machinelearningmastery.com/what-is-imbalanced-classification/

Brownlee, J. (2021, January 4). *Random oversampling and undersampling for imbalanced classification*. Machine Learning Mastery. Retrieved November 3, 2021, from https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/

Rdrr.IO (n.d.), Retrieved November 3, 2021, from https://rdrr.io/cran/performance/man/check_collinearity.html

Sunrisers FAQ. M1. (n.d.). Retrieved November 3, 2021, from https://www.m1.com.sg/Support/FAQ/sunrisers-faq

Bock, T. (2021, June 9). *Decision trees are usually better than logistic regression*. Displayr. Retrieved November 3, 2021, from https://www.displayr.com/decision-trees-are-usually-better-than-logistic-regression/

Nemes, A. (2021, September 15). *Telecommunication Loyalty Programs: A comprehensive guide (2021)*. Antavo. Retrieved November 3, 2021, from https://antavo.com/blog/telecommunication-loyalty-programs/

Benedict, C. (2020, June 12). *How to improve customer retention in the telecom industry*. Acquire. Retrieved November 3, 2021, from https://acquire.io/blog/improve-customer-retention-in-telecom-industry/

Llopis, G. (2016, August 9). *Consumers are no longer brand loyal*. Forbes. Retrieved November 3, 2021, from https://www.forbes.com/sites/glennllopis/2014/12/10/consumers-are-no-longer-brand-loyal/?sh=179b05792ae0

Amy, G. (2014, November 2). *How valuable are your customers?* Harvard Business Review. Retrieved October 22, 2021, from https://hbr.org/2014/07/how-valuable-are-your-customers

How telcos can raise the game with their bundling strategy. Apigate. (2021, September 13). Retrieved November 3, 2021, from https://www.apigate.com/how-telcos-can-raise-the-game-with-their-bundling-strategy/

Vdovjak, D. (2021, February 9). *The future of Telco Bundling – New Trends and challenges*. LinkedIn. Retrieved November 3, 2021, from https://www.linkedin.com/pulse/future-telco-bundling-new-trends-challenges-danijela-vdovjak

Laura, M. (2021, August 26). *Product bundling in Telecommunications*. Acro Media. Retrieved November 3, 2021, from https://blog.acromedia.com/product-bundling-in-telecommunications